Basira

Document Intelligence Platform

—

This documentation describes an automated workflow for ingesting, processing, and validating documents for **Basira** using Google Cloud services.

Services Used

Cloud Storage: It acts as the landing zone for incoming documents (**Source bucket**) and the holding area for files needing manual review (**Validation bucket**).

Cloud Functions: To trigger the processing pipeline as soon as a new document arrives in the source bucket.

Document AI: As it has a built-in OCR feature, it will be a good fit to extract information from the documents.

I believe that the incoming documents, such as **Bank statements** and **Invoices,** will be in a **structured format** and as **Text-based** documents, and this will be very helpful to increase the confidence score.

BigQuery: To store the final results in a structured format for analytics in BigQuery and ML using BigQuery ML.

Architecture Flow

Document Ingestion

Documents are uploaded from an external source via an API.

The files land in the **Cloud Storage** (Source bucket). This bucket is the starting point for all incoming documents.

Trade-off: Receiving documents via an API to Cloud Storage is simple and scalable, but it depends entirely on the external system's reliability. If the API sends corrupted files or duplicates, the pipeline will not produce the required results.

Process Triggering

Cloud Function is set up to monitor the Source bucket. Upon the arrival of a new document, this function is automatically triggered, initiating the processing pipeline, which begins with the Document AI service.

Trade-off: Using Cloud Functions keeps the flow serverless and **cheap**, but functions have cold starts and timeout limits, and this might delay processing.

Data Extraction (OCR)

The triggered Cloud Function sends the new document to the **Document AI** service.

Document AI processes the file, extracts the relevant text and data, and converts it into a structured format.

Trade-off:

I believe that the incoming documents, such as **Bank statements** and **Invoices,** will be in a **structured layout** and as **Text-based** documents, and this will be very helpful to increase the confidence score.

So, performance drops for scans, images, or noisy layouts. Relying heavily on text-based documents means the pipeline may struggle when less clean documents appear.

Confidence Score (Validation)

Document AI provides a **confidence score** for the accuracy of its extraction. This workflow uses a **95% threshold** to route the data.

- **Path A (High Confidence)**: If the score is > 95%, the data is considered accurate. It is loaded in its structured format directly into **BigQuery** for analysis and ML models.
- **Path B (Low Confidence)**: If the score is < 95%, the document is flagged for manual review. It is automatically moved to a separate Cloud Storage (Validation bucket) to await human-in-the-loop (HITL) correction.

  The human-in-the-loop (HITL) is a mandatory part of the process to enhance the way that Document AI extracts the data, and the OCR will be enhanced.

  With time, the Document AI will be much smarter and confident while reading the documents

Trade-offs:

- **Confidence Score Logic**
  Choosing the wrong threshold can either overwhelm manual review or let low-quality data into BigQuery.

- **High Confidence in BigQuery**
  Sending high-confidence data straight to BigQuery is fast and cheap, and it can be minimized by applying a clear schema for the target tables; still, it risks ingesting extraction errors.
- **Low Confidence to HITL Review**
  Human review increases accuracy and improves future models, but it adds cost, latency, and operational overhead. If many documents fall below the threshold, the process becomes slow and expensive.